Intelligibility of dialects and related languages

›Charlotte Gooskens

LOT course Tilburg, 13 January 2012

Introduction

Project I

Linguistic determinants of mutual intelligibility in Scandinavia

- VIDI, financed by NWO
- 1 January 2006 – 1 June 2011

Project II

Mutual intelligibility of language varieties in the Low Countries: linguistic and attitudinal determinants

- VNC, financed by NWO and FWO
- 1 January 2007 – 1 January 2011
- Project members from Leuven, Nijmegen and Groningen

Project III

Mutual intelligibility of closely related languages in Europe: linguistic and non-linguistic determinants

- NWO Free Competition
- 1 September 2011 – 1 September 2016

Intelligibility of dialects and related languages

Intelligibility:

- The degree to which a speaker of one variety understands the speech of another closely related variety
- Can be expressed in a single number

Dialects and related languages:

- Distances can be expressed in a single number

Intelligibility of dialects and related languages

Assumptions:

- First confrontation (inherent intelligibility)
- Spoken language only

Intelligibility of dialects and related languages

Similarities to:

- defective speech
- speech in noise
- foreign accents
- talking machines

Intelligibility of dialects and related languages

Mutual intelligibility:

- Haugen (1966): semicommunication
- ≈ nonconvergent/asymmetric/bilingual discourse, receptive bilingualism
- Speakers of different but related languages each speak their own language and still comprehend one another's languages
- Mutual intelligibility is sometimes imperfect and asymmetric

Semicommunication

Prerequisites:

- Language community
- Interaction
- Symbolic integration

Semicommunication

Observed semicommunication (Zeevaert 2004):

- Danish - Norwegian - Swedish (Haugen 1966, Maurud 1976….)
- Czech - Slovakian (Budovičá 1987)
- Czech - Polish (Hansen 1987)
- Spanish - Portuguese (Coseriu 1988, Jensen 1989, Zeevaert 2002)
- Italian - Spanish (Hansen 1987)
- German - Dutch (Haz 2002)
- Frisian - Dutch (Feitsma 1986)
- Croatian - Serbian (Haugen 1990)
- Hindi - Urdu (Haugen 1990)
- Icelandic - Faeroese (Braunmuller & Zeevaert 2001)
- Macedonian - Bulgarian (Haugen 1990)
- Russian - Bulgarian (Braunmuller & Zeevaert 2001)
- Chinese dialects (Cheng 1997, Tang & Van Heuven 2007)
- Arabian dialects (Haugen 1990)

Factors explaining intelligibility

Extra-linguistic

- attitude
- contact
- linguistic experience
- orthography

Linguistic

- **sounds**
- prosody
- **lexicon**
- morphology
- syntax

Factors explaining intelligibility

- A model of intelligibility: the relative importance of the factors

- Intelligibility measurements can be used to find out how the linguistic factors should be weighed

Central questions

1. How can the mutual intelligibility between closely related languages be measured?

2. How can the relevant (extra-)linguistic factors be measured?

3. To what extent are the (extra-)linguistic factors predictors of intelligibility?

Overview

1. intelligibility testing
2. relationship between intelligibility and phonetic distances
3. relationship between intelligibility and lexical distances
4. conclusion

Measuring intelligibility

Measuring intelligibility

- **Opinion testing:**
  How well does the listener **think** he understands the other language variety?

- **Functional testing:**
  How well does the listener **actually** understand the other language variety?

- **Observations:**
  How well do people understand each other in **real** language situations?

Opinion testing

How well does the listener **think** he understands the other language variety (opinion scores)?

Advantages:

- efficient
- the same words can be tested in each variety

Disadvantages:

- listeners may not be able to judge intelligibility

Functional testing

How well does the listener **actually** understand the other language variety?

Advantages:

- actually measures intelligibility

Disadvantages:

- priming effects must be avoided
- heavy memory load
- time consuming

Functional testing

Text intelligibilty:

- tests language as a whole
- resembles a natural situation

Word intelligibility:

- gives researcher the opportunity to investigate the role of specific linguistic factors
- artifical situation

Measuring intelligibility

Methods:

- open questions
- multiple choice
- cloze test
- translations
- opinion scales

observations

Observations of real language situations:

- Number of misunderstandings, repairments, reformulations, pauses, turn taking etc.
- Arranged or real conversations

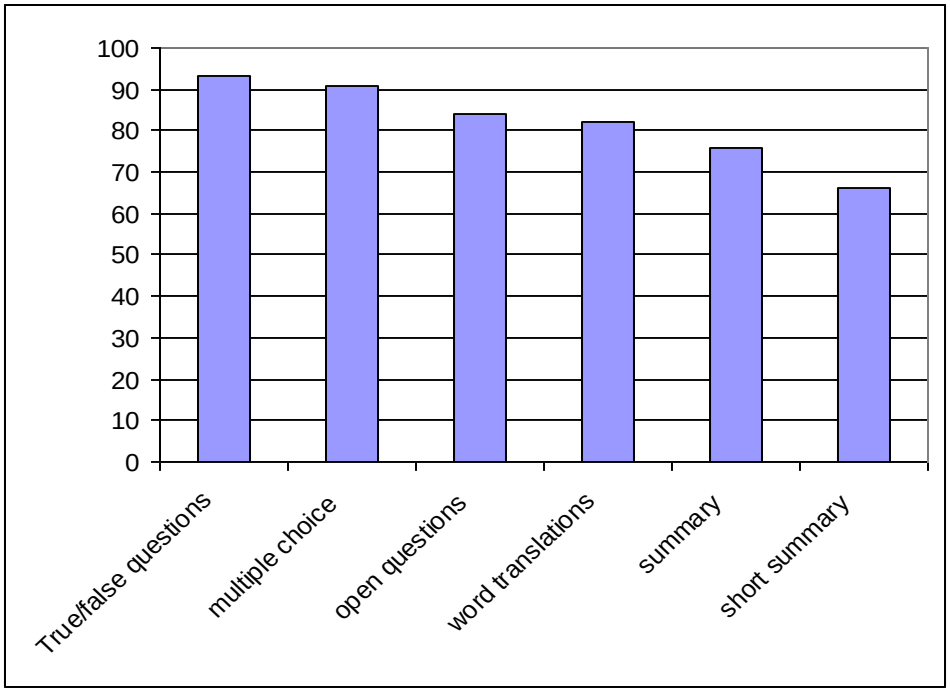
Comparison of methods

The intelligibility of the same Swedish text by Danes tested in **six different test conditions** (Doetjes 2007):

1. Open questions
2. True/false questions
3. Multiple choice questions
4. Word translation
5. Summary
6. Short summary

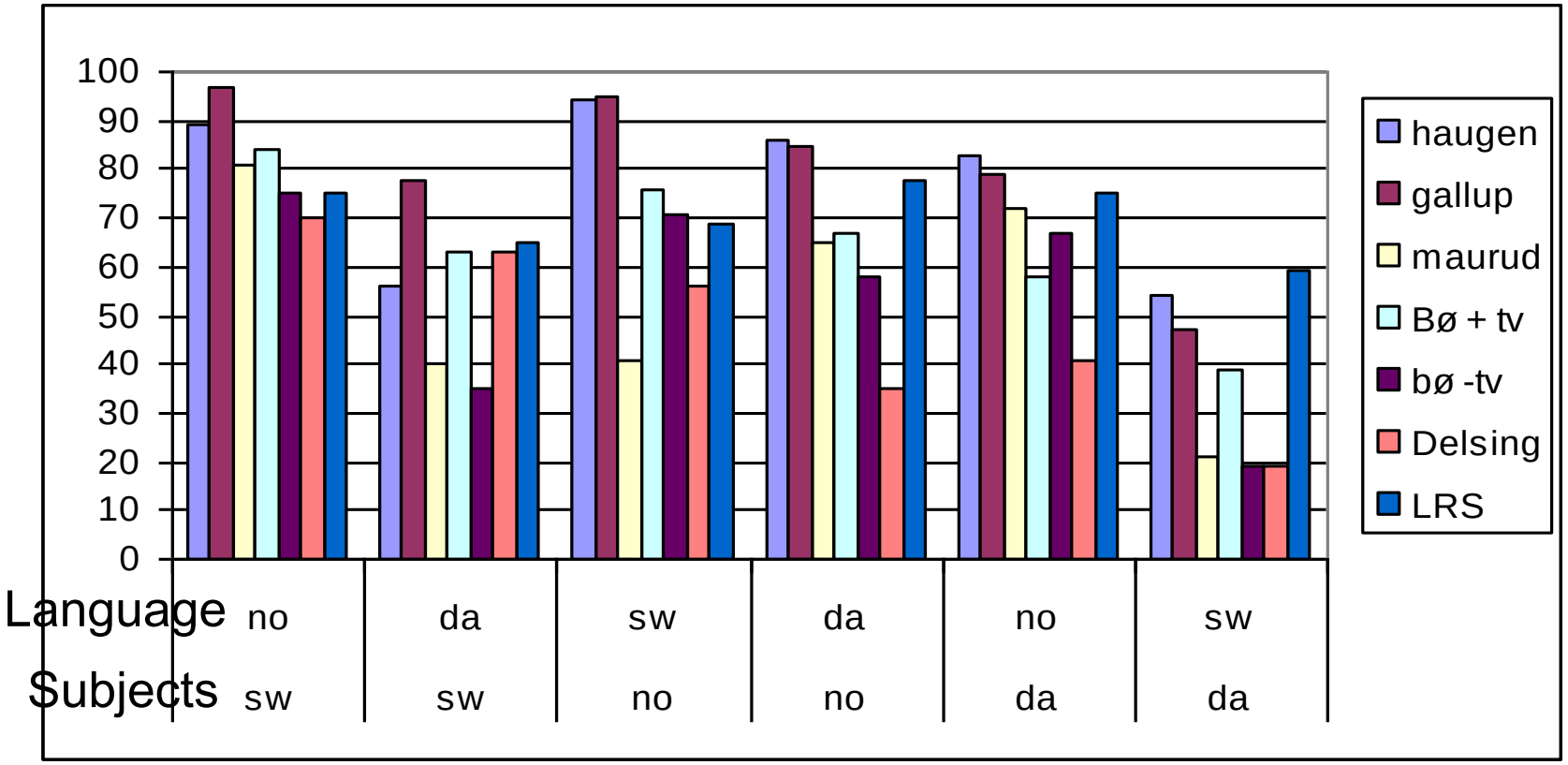
Comparison of methods

The intelligibility of the same Swedish text by Danes tested in **six different test conditions** (Doetjes 2007):



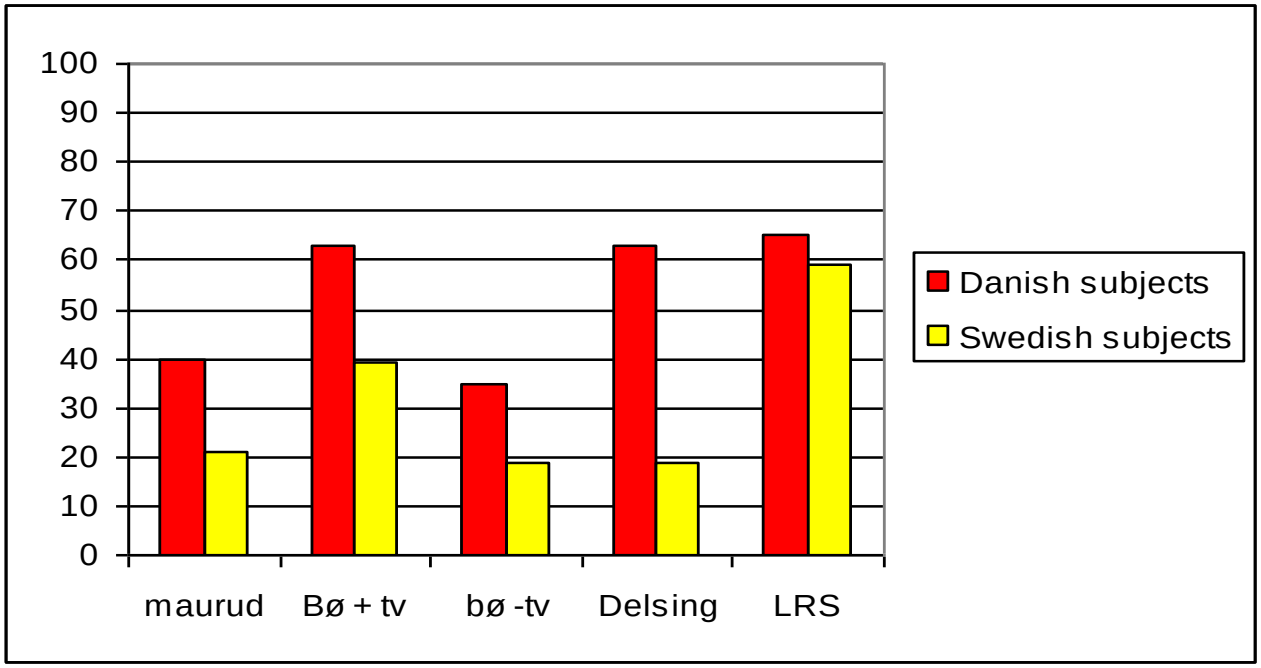
Comparison of methods

All Scandinavian investigations:



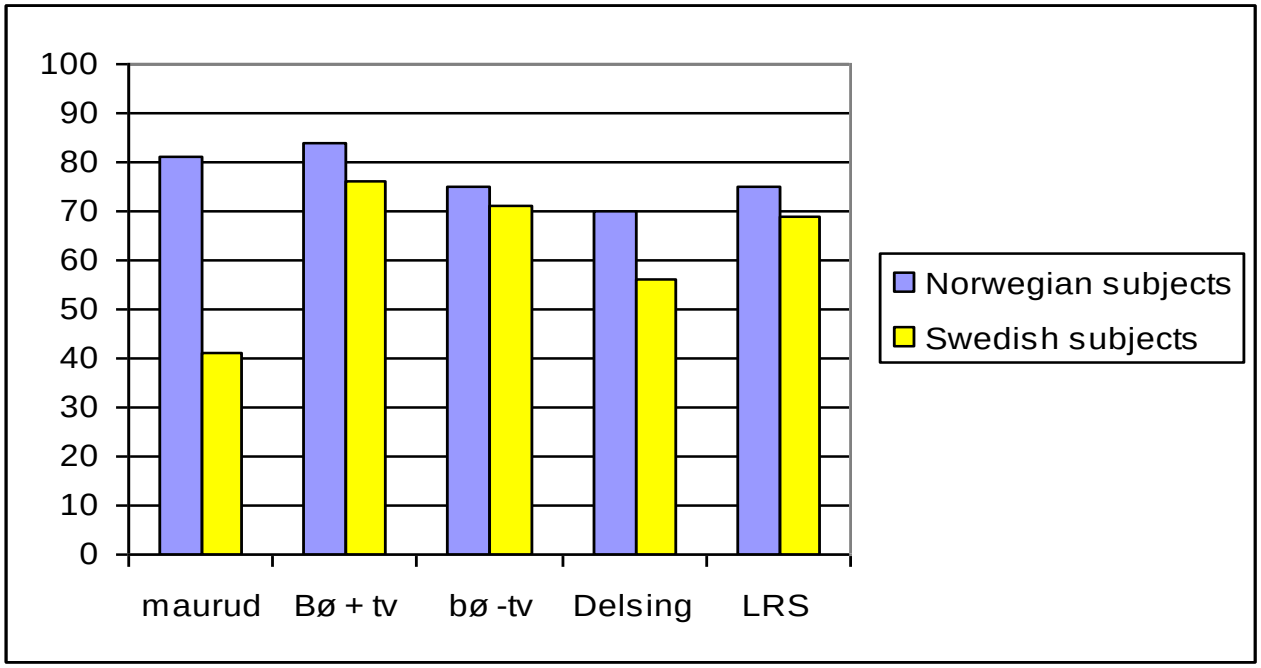
Comparison of methods

Swedish – Danish mutual comprehension



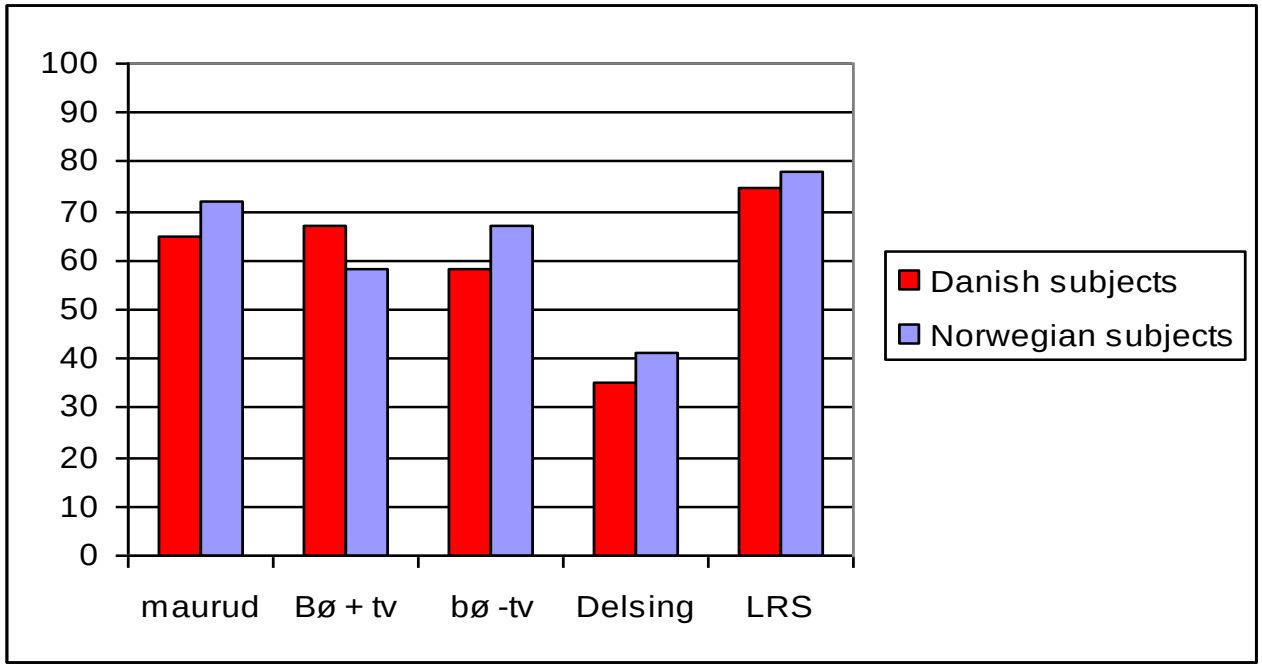
Comparison of methods

Swedish – Norwegian mutual comprehension



Comparison of methods

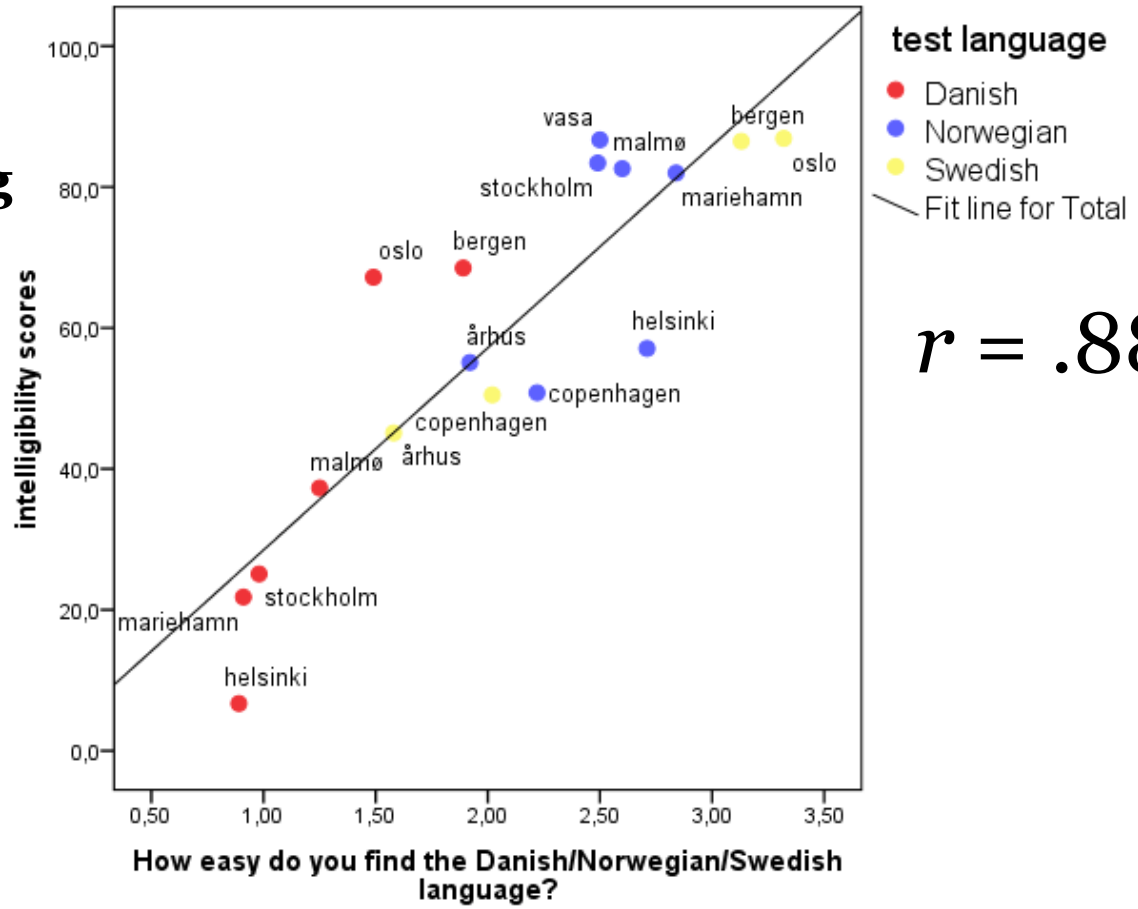
Danish – Norwegian mutual comprehension



Comparison of methods

Opinion versus functional testing

Delsing & Lundin Åkesson (2003):



$r = .88$

Factors explaining intelligibility

Extra-linguistic

- attitude
- contact
- linguistic experience
- orthography

Linguistic

- **sounds**
- prosody
- lexicon
- morphology
- syntax

Relating Levenshtein distances to intelligibility scores

Research question

- How well can Levenshtein distances predict intelligibility?

Investigations

Beijering, Gooskens & Heeringa (2008):

The intelligibility of 18 Nordic language varieties among Danes

Kürschner, Gooskens & Van Bezooijen (2008):

The intelligibility of Swedish words among Danes

Gooskens, Van Bezooijen & Van Heuven (in press):

The mutual intelligibility of 100 words among Dutchmen and Germans

The intelligibility of 18 Nordic language varieties among Danes

Material

- recordings of ‘The North Wind and the Sun’
- 18 Nordic language varieties
- mean 98 words
- phonetic transcriptions of the cognates (historically related words)

Material

The North Wind and the Sun *were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.*

Material


Tromsø
Torshavn
Rana
Bjugn
Storliden
Oppdal
Gaular
Trysil
Helsinki
Oslo
Fyresdal
Stockholm
Lidköping
Høgsted
Gryttinge
Katrad
Hjordkær
Lyngby

Intelligibility

Stimulus material:

› 6 sentences in 6 varieties (latin square design)

Test persons:

› 18 groups of high school pupils from Copenhagen, aged between 15 and 20 (average 17.6)

Task:

› translate word for word into Standard Danish

Correlation

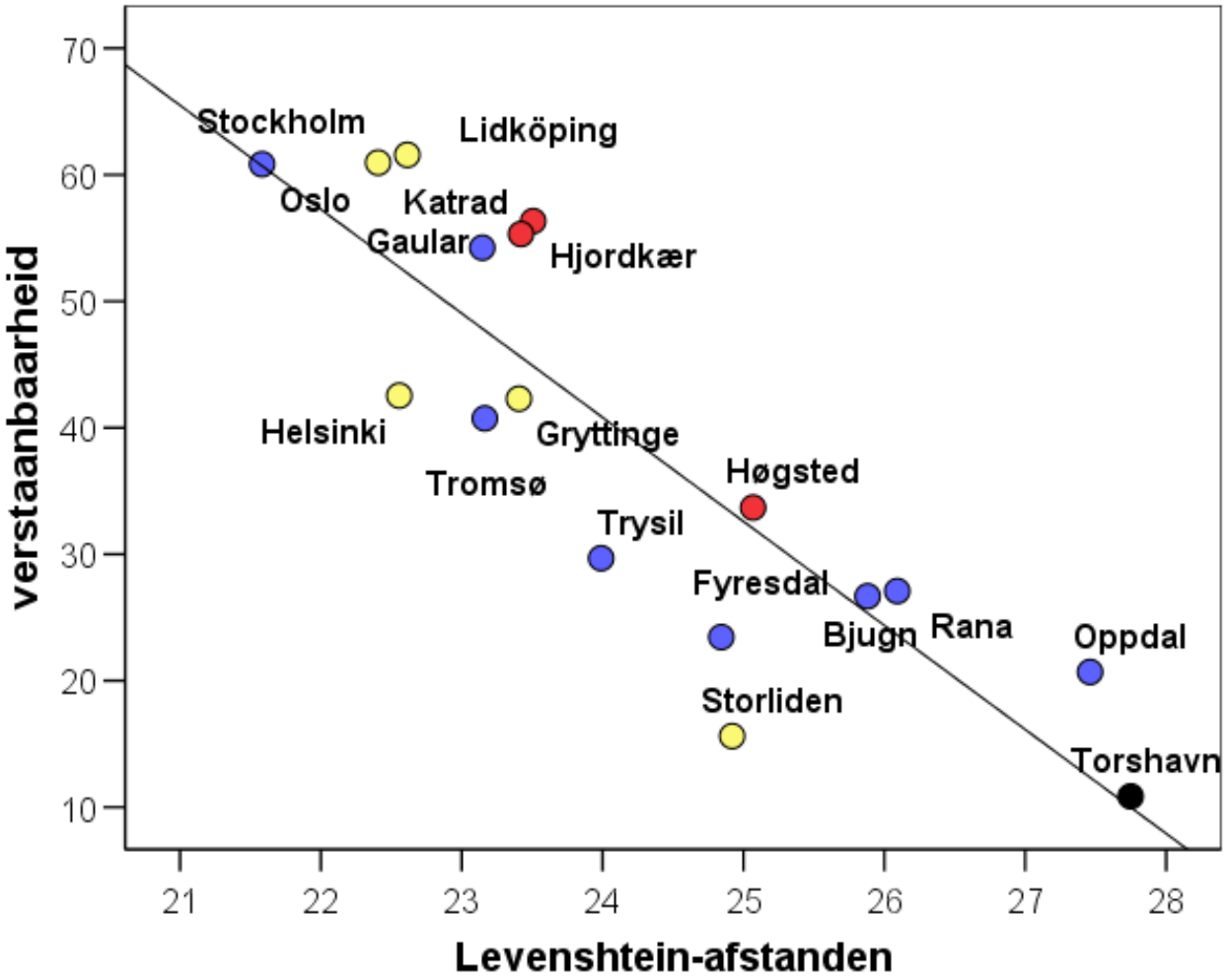
Dependent variable :

› % correctly translated words per variety

Independent variable:

› Levenshtein distances between Standard Danish and the 18 varieties

Relation between intelligibility and Levenshtein distances



r=-.86

Relation between intelligibility and Levenshtein distances

- It is possible to predict intelligibility of language varieties to a high extent by means of Levenshtein distances

Consonants vs vowels

- Are consonants or vowels more important for the intelligibility?

Consonants vs vowels

- Word recognition in English depends more on correct **consonant** identification than on the correct identification of **vowels** (Van Ooijen 1994 and references).

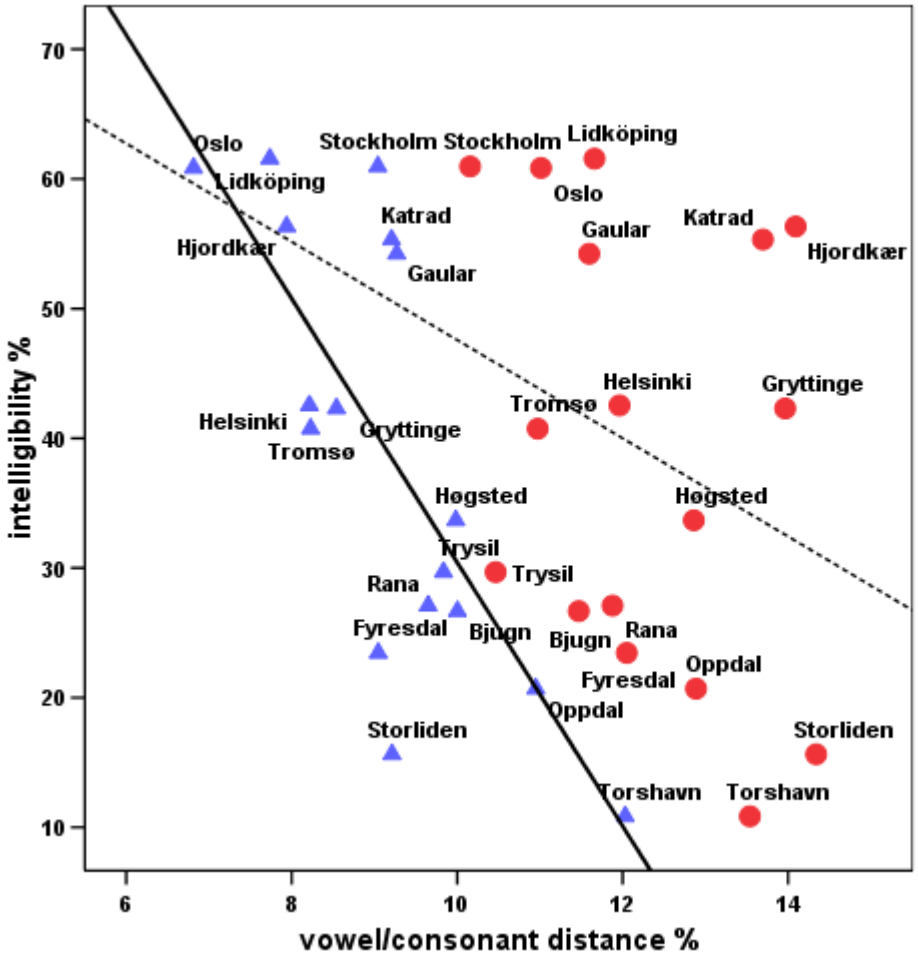
- **Consonants** are more important for the semantic identity of a word than **vowels**; they function as reference points in words.

Consonants vs vowels

Hypothesis:

› Deviations in vowels are less damaging for the intelligibility of a closely related language than deviations in the consonants.

Relation between intelligibility and Levenshtein distances, vowels and consonants



vowels

r=-.29

consonants

r=-.74

The intelligibility of Swedish words among Danes

Word intelligibility

Material:

- Recordings of 347 frequent Swedish cognates
- Phonetic transcriptions

Test persons:

- 38 Danish high school pupils aged 16-19

Task:

- Translate words into Standard Danish

Word intelligibility

Independent variable:

› Levenshtein distances between Swedish and Danish words

Dependent variable :

› % correct translations per word

Word intelligibility

Relation between intelligibility and Levenshtein distances:

- r=-.27
- Other factors considered in order to predict intelligibility at the word level

Word intelligibility

11 factors considered for prediction of intelligibility of Swedish words by Danish listeners:

- **Levenshtein distance**
- **Foreign sounds**
- **Word length**
- Word stress differences
- **Differences in number of syllables**
- **Neighbourhood density**
- Lexical tones
- ***Stød***
- Etymology (native words versus loan words)
- **Orthography**
- **Word frequency**

Word intelligibility

Correlation still low:

- Low correlations for all factors
- Logistic regression: $R^2 = .21$

Why?

- Logistic regression model
- Idiosyncraties of individual words?

Word intelligibility

Asymmetry:

- Danes translated 61% of the Swedish words correctly
- Swedes translated 49% of the Danish words correctly

Why?

- Different speakers
- Different backgrounds of listeners
- Different language attitudes
- Linguistic factors?

The mutual intelligibility of 100 words among Dutchmen and Germans

Word intelligibility

Subjects:

- 63 Dutchmen and 56 Germans
- 9 - 12 years
- no previous knowledge of test language
- equally positive attitudes

Stimuli:

- 100 frequent cognate nouns
- perfect bilingual speaker

Task:

- translation

Word intelligibility

Results:

- Dutch children 50% correct translations
- German children 42% correct translations
- Difference is significant at .001 level

- Correlation with Levenshtein distances: r=.46 for both groups

Word intelligibility

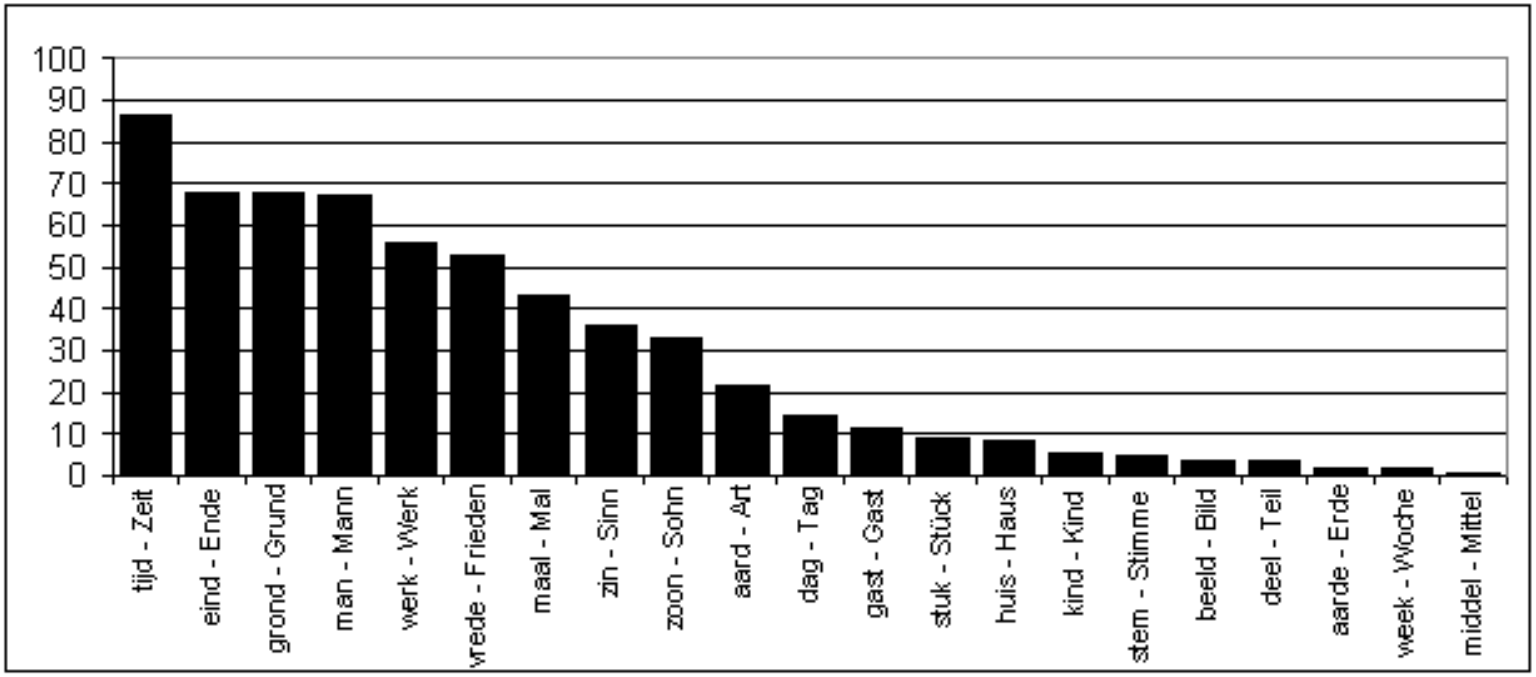
Questions:

- Why is correlation between intelligibility and Levenshtein distances low?
- Why is intelligibility asymmetric?

Analysis of errors can give information about listener strategies

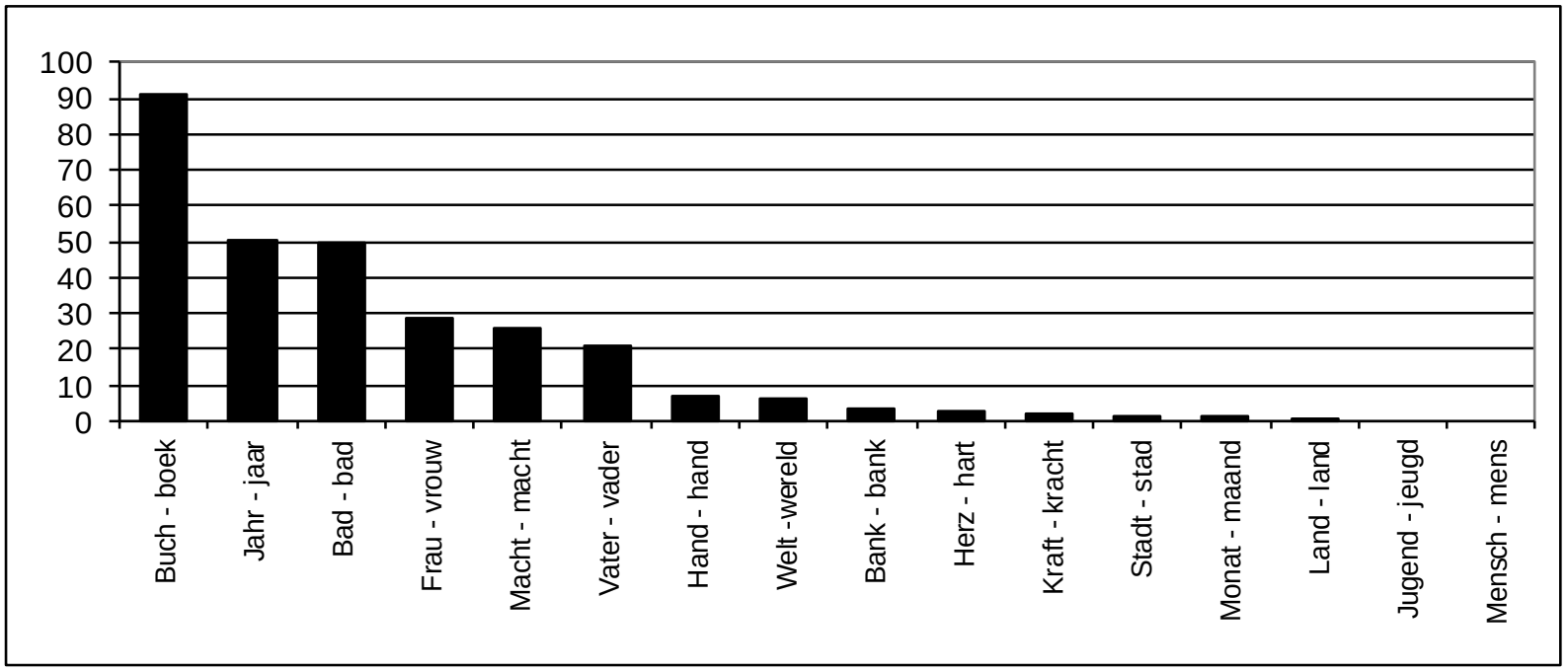
Word intelligibility

Cognates that were better understood by Dutch subjects than by German subjects:



Word intelligibility

Cognates that were better understood by German subjects than by Dutch subjects:



Word intelligibility

- Phonetic confusions of sounds:

  e.g. Du. /x/ is often perceived as /h/ by Germans

  Du. *grond,* Ge. *Grund* ‘ground’ is translated into Ge. *hund* ‘dog’

Word intelligibility

- Subtle phonetic differences not present in phonetic transcription:

  e.g. Du. /l/ is often velarized and perceived as a diphthong Germans
  Du. *maal,* Ge. *Mal* ‘time’ is translated into
  Ge. *maus* ‘mouse’

Word intelligibility

Perception of a sound may depend on position in word:

- Du. pre-consonantal /r/ is problematic: Du. *werk,* Ge. *Werk* ‘work’ is translated into Ge. *Weg* ‘road’
- But in word final position Ge. /r/ is problematic: Ge. *Jahr*, Du. *jaar* is translated into Du. *ja* ‘yes’

Word intelligibility

› Influence of neighbour words:

e.g. Ge. *Bad*, Du. *bad* ‘bed’ is often translated into Du. *paard* ‘horse’ or *baard* ‘beard’

– no alternatives in German

Word intelligibility

› Interference from foreign languages:

e.g. Du. *tijd*, Ge. *Zeit* 'time' is often translated into English *date* or *dad* by Germans

Word intelligibility

- Phonetic details may play an important role in the intelligibility of cognates in unpredictable ways
- Phonotactic constraints may effect the perception of sounds in some positions but not in others
- Presence of lexical neighbours and false friends may influence word recognition

Relating lexical distances to intelligibility scores

Lexical distances

Simple measure:

% cognates in a representative language sample (text, corpus, frequency word list etc.)

Cognate:

e.g. English *butter* Dutch *boter*

Non-cognates:

e.g. English *butter* Danish *smør*

Lexical distances

Can be **asymmetric**:

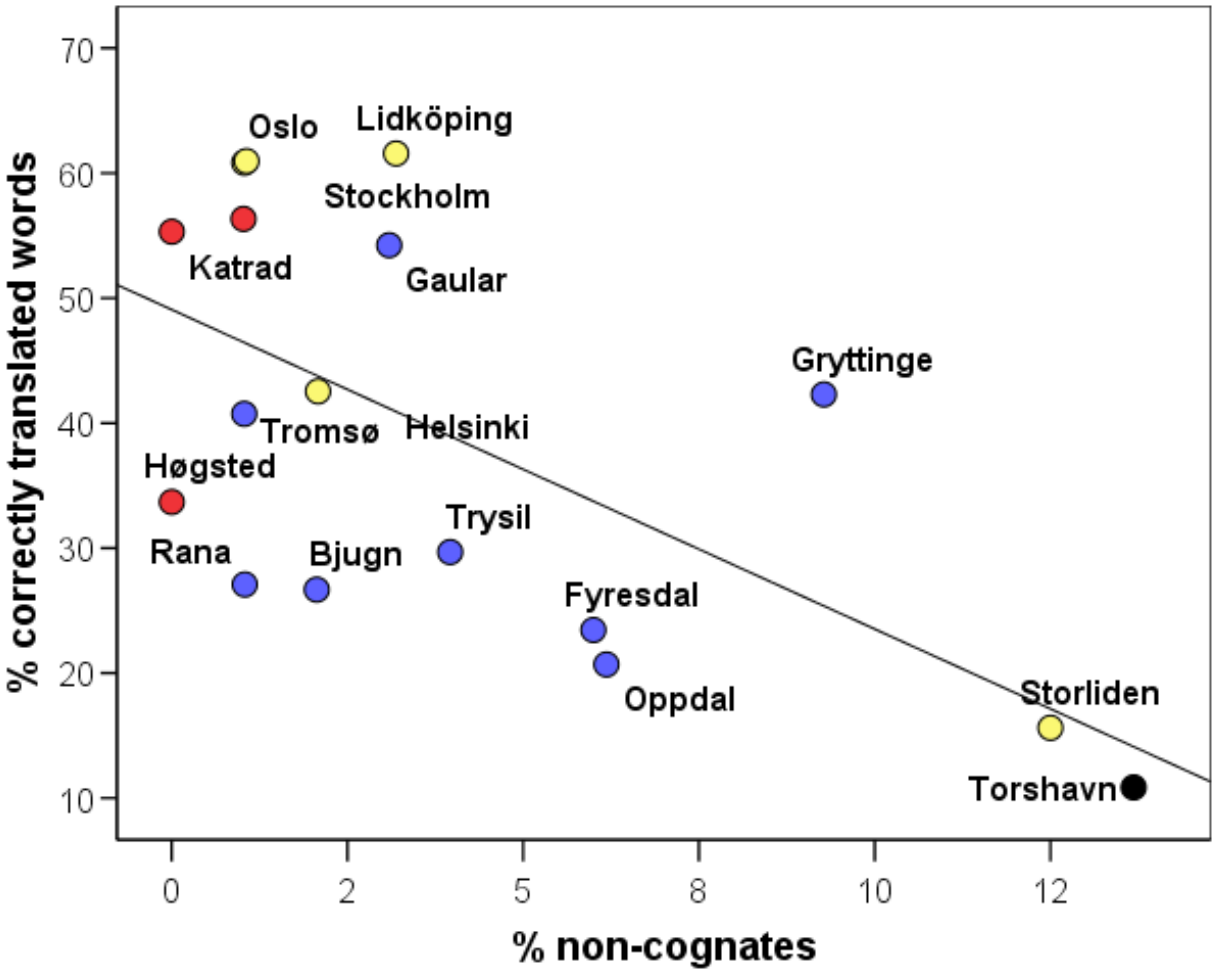
Example ‘room’

Swedish	Danish
rum	*rum*
	værelse

Lexical distances should be measured from A to B and from B to A

Lexical distances

Lexical distances



r = -.64

Lexical distances

To be considered:

- Speaking style
- Lexical domain
- Word class
- Content word vs. function word
- Frequency
- …

Lexical distances

Stepwise lineair regression analysis with the independent variables phonetic distance and lexical distance:

R=-.86

Lexical distances do not contribute significantly

Conclusions

1. Intelligibility can to a high extent be predicted by phonetic distances, but…

2. …phonetic details may play an important role in the intelligibility of cognates in unpredictable ways

3. Lexical distances play a smaller role in the intelligibility of Scandinavian language varieties, but…

4. …their role may be language dependent